



Stroke analysis and recognition in functional near-infrared spectroscopy signals using machine learning methods

TIANXIN GAO,¹ SHUAI LIU,¹ XIA WANG,² JINGMING LIU,² YUE LI,³
XIAOYING TANG,¹ WEI GUO,^{2,5} CONG HAN,^{4,6} AND YINGWEI
FAN^{1,7} 

¹School of Medical Technology, Beijing Institute of Technology, 100081, Beijing, China

²Beijing Tiantan Hospital, Capital Medical University, 100050, Beijing, China

³Department of Biomedical Engineering, School of Medicine, Tsinghua University, 100084, Beijing, China

⁴Department of neurosurgery, the Fifth Medical Center of PLA General Hospital, 100071, Beijing, China

⁵guowei1010@126.com

⁶hc82225@126.com

⁷fanyingwei@bit.edu.cn

Abstract: Stroke is a high-incidence disease with high disability and mortality rates. It is a serious public health problem worldwide. Shortened onset-to-image time is very important for the diagnosis and treatment of stroke. Functional near-infrared spectroscopy (fNIRS) is a noninvasive monitoring tool with real-time, noninvasive, and convenient features. In this study, we propose an automatic classification framework based on cerebral oxygen saturation signals to identify patients with hemorrhagic stroke, patients with ischemic stroke, and normal subjects. The reflected fNIRS signals were used to detect the cerebral oxygen saturation and the relative value of oxygen and deoxyhemoglobin concentrations of the left and right frontal lobes. The wavelet time-frequency analysis-based features from these signals were extracted. Such features were used to analyze the differences in cerebral oxygen saturation signals among different types of stroke patients and healthy humans and were selected to train the machine learning models. Furthermore, an important analysis of the features was performed. The accuracy of the models trained was greater than 85%, and the accuracy of the models after data augmentation was greater than 90%, which is of great significance in distinguishing patients with hemorrhagic stroke or ischemic stroke. This framework has the potential to shorten the onset-to-diagnosis time of stroke.

© 2023 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

1. Introduction

Stroke is a clinical emergency caused by cerebral infarction or hemorrhage caused by blockage or rupture of blood vessels [1]. Worldwide, stroke is the second most common cause of death and the third most common cause of disability [2]. In 2020, the global death toll of cerebrovascular diseases was 7.08 million (including 3.48 million deaths from ischemic stroke, 3.25 million deaths from intracerebral hemorrhage (ICH), and 0.35 million deaths from subarachnoid hemorrhage) [3]. Reducing the burden of stroke requires intervention across the health system from primary prevention through diagnosis, acute treatment, rehabilitation and secondary prevention [4]. Timely diagnosis and treatment can save lives and prevent severe disabilities [5].

The acute management of stroke patients requires a fast and efficient screening imaging modality. Computed tomography (CT) is the most widely used diagnostic tool for stroke because of its availability and rapid acquisition time [6]. The imbalance in the distribution of specialized stroke medical resources between urban and rural areas has led to a higher burden of stroke in rural areas than in urban areas [7]. The onset-to-imaging time of CT was 107 minutes

and the onset-to-angiography time was 213 minutes in a study from France [6]. However, the onset-to-door time might be more than 24 hours in some rural areas [5]. For rural areas without equipment such as CT or for patients who cannot reach the emergency room in a short time, portable equipment might shorten the diagnosis time of stroke, hence saving lives and reducing disabilities caused by stroke. Even in urban areas, portable screening equipment in the ambulance could save onset-to-needle time by sending the patient directly to the proper hospital [8].

Functional near infrared spectroscopy (fNIRS) provides simple, continuous, real-time and noninvasive brain monitoring. Portable fNIRS equipment has been used in stroke studies [9–17], with some limitations [18]. To address the limitations of detection depth and obtain more physiological information from the signal, our previous studies have found that spontaneous low-frequency oscillations (LFOs) of cerebral oxygenation signals by fNIRS showed significant differences between cerebral infarction and nonstroke patients [19]. Spontaneous LFOs could reflect the hemodynamic changes of whole cerebral blood vessels, which are not restricted to the detection area. Other studies have also confirmed this phenomenon [20]. LFOs in functional imaging data have gained increased interest in the study of disease caused cognitive decline [21], cerebral autoregulation disorder [22], headache [23], the covert consciousness in neurocritical patients [24], as well as in functional study of normal brain [25–27]. This study proposes a new clinical usage of LFOs based on the absolute value of parameter obtained from the spatially resolved spectroscopy algorithm. Machine learning techniques have been used to analyze the association and rank the importance of the risk factors for stroke [28]. The most popular machine learning model applied for stroke risk prediction is the support vector machine (SVM) [18]. Hence, to further analyze the characteristics of patients with hemorrhagic stroke, and ischemic stroke and normal individuals, we conducted this study. Here, machine learning models based on spontaneous LFO signals were used to help identify the characteristics of different types of strokes.

2. Methods

2.1. Subjects

The study selected 113 patients with acute cerebral hemorrhage or cerebral infarction from the emergency department of Beijing Tiantan Hospital affiliated with Capital Medical University and the Department of Neurosurgery, the fifth medical center of PLA General Hospital, and there were also 64 healthy control group participants. The data were collected within two weeks after the onset of stroke. After excluding patients who were older than 80 or younger than 30, 40 patients with cerebral hemorrhage, 40 patients with cerebral infarction and 40 healthy subjects were enrolled. The corresponding quantity for each category is shown in Table 1. This study was approved by the Fifth Medical Center of the PLA General Hospital Research Ethics Committee (No. KY-2022-9-69-1).

Table 1. The data scale and categories

Categories	Number of participants	Quantity used
Control	64	40
Hemorrhagic stroke	51	40
Ischemic stroke	62	40

2.2. Near infrared signal measurement

The EGOS-600A (EnginMed, Co., Ltd. Suzhou, China) [29–35] is a CW (continuous wave) recording system with four channels. Each channel is connected to a reflective probe. Each probe has 1 source and 2 detectors, with source-detector distances of 3 cm and 4 cm. The probe's light

source has three wavelengths (760 nm, 810 nm, and 840 nm). The spatially resolved spectroscopy (SRS) algorithm [35] was used to calculate tissue oxygen saturation (TOI, also known as regional cerebral oxygen saturation and local oxygen saturation). The modified Lambert Beer law was used to obtain the relative values of oxygen and deoxyhemoglobin concentrations (HbO, Hb) through the raw fNIRS signal.

For this study, two channels were used to record the cortical blood oxygen changes from the subjects enrolled at a sampling rate of 10 Hz. The two probes were fixed on the left and right sides of the forehead (as shown in Fig. 1). According to the international 10–20 electrode system, two pairs of light detectors were placed at positions Fp1-F7 and Fp2-F8, and corresponding light sources (the red blocks on the probe in Fig. 1) were placed at F7 and F8, 3 and 4 cm from the two detectors (the blue blocks on the probe in Fig. 1), respectively.

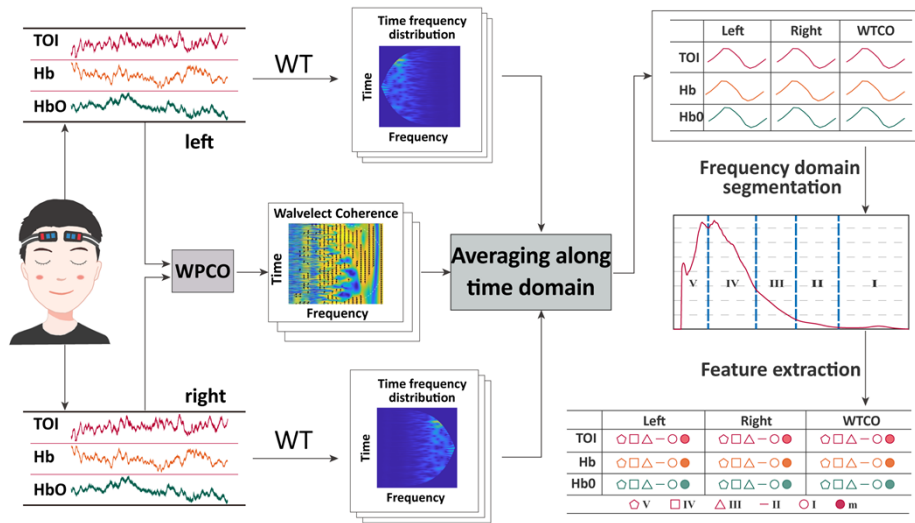


Fig. 1. Research flowchart. The figure shows the wavelet transform for calculating the cerebral oxygen signals of each side and the wavelet phase coherence between the left and right cerebral oxygen signals of each subject. The frequency domain signals are obtained by averaging the two-dimensional time-frequency results in the time domain, and then the signals of different frequency bands are extracted from the frequency signals to obtain the characteristics. (6 features extracted from each of the 9 frequency signals, 54 for each subject).

None of the subjects had smoked, drank or taken drugs within 24 hours before the experiment. During the experiment the subjects closed their eyes and remained still. The room was dark, and the room temperature was 25 °C. The left and right frontal lobe oxygenation signals were monitored continuously for 20 minutes. Figure 1 shows the whole experimental process.

2.3. Time frequency feature analysis

For the time-domain signal we obtained, we use the wavelet transform method to conduct time-frequency transformation and analyze the signal in the frequency domain. MATLAB 2018 (MathWorks, MA, USA) was used to conduct time-frequency analysis on each parameter. A one-dimensional frequency domain signal curve was obtained by averaging the two-dimensional wavelet amplitude matrix along the time dimension. Each frequency signal was divided into five physiological frequency bands according to the physiological cause of the oscillation. As shown in Fig. 1, the five bands are represented by Roman numerals I to V, which denote cardiac activity (0.4000 to 2.0000 Hz), respiratory activity (0.1500-0.4000 Hz), myogenic activity

(0.0600-0.1500 Hz), neurogenic activity (0.0200-0.0600 Hz) and endothelial metabolic activity (0.0095 to 0.0200 Hz) [19,36–48], respectively. The average value of each frequency band was obtained as the feature parameters of the frequency band. The average value of each time series signal was also extracted as a feature. Eighteen features of unilateral cerebral vessels of a subject were extracted.

Based on the wavelet transform, wavelet phase coherence (WPCO) can reveal the possible relationship between two signals by evaluating their instantaneous phase matching. The WPCO value is calculated according to the frequency domain amplitude of the instantaneous phase difference, and the average value is calculated. Its value ranges from 0 to 1, and it can be used to evaluate the phase coherence between two signals at the same frequency and time. In a previous study [19], WPCO analysis was used to explore the relationship with blood oxygen saturation within a specific frequency range and to research the connectivity of left and right brain functions.

The phase synchronization between left and right prefrontal oscillations is likely related to cognitive function. In this study, the WPCO method was also used to evaluate the phase coherence of cerebral oxygen signals in the left and right prefrontal tissues of patients with cerebral hemorrhage, patients with cerebral infarction and the control group at rest. By reflecting the synchronization between the instantaneous phases of signals measured in the continuous process of the left and right frontal brain regions, this synchronization information can be used to evaluate the functional relationship of different cerebral cortex regions. This evaluation provides a new understanding of the dynamic regulation of brain function from the perspective of phase. WPCO is calculated for each pair of time series signals. In addition, the frequency sequences obtained from each of the five frequency bands were averaged, and each WPCO signal was averaged in the entire frequency domain. Therefore, 18 features were extracted based on the connection between the two cerebral hemispheres. The above features are represented by the symbols in Table 2, where “cm” represents the calculation method and “L” and “R” represent the WT and the average value in the time domain of the left or right frontal lobe signal, respectively. “w” represents WPCO of the left and right frontal lobe signals. “Par” indicates the parameter: TOI, HbO or Hb. “Fb” indicates the frequency band or frequency interval in which the frequency signal is averaged, as shown in Table 2. “m” represents the average value of the time series signal for WT and the average value of the frequency signal in the whole frequency domain for WPCO. For example, “RTOI_II” represents the II frequency band feature of the right frontal lobe TOI signal.

Table 2. Feature symbols and meanings of each part

cmPar_fb		
cm	Par	fb
L, R, w	TOI, HbO, Hb	V, IV, III, II, I, m

One-way ANOVA and the t test were used to evaluate the significant differences in features among patients with cerebral hemorrhage, patients with cerebral infarction and controls. Table 3 shows the statistical results of the features of patients with cerebral hemorrhage, patients with cerebral infarction and healthy controls ($\diamond P < 0.05$).

2.4. Difference analysis of cerebral oxygen signals in patients with cerebral hemorrhage and cerebral infarction based on machine learning

For each parameter (TOI, HbO, or Hb), we trained and built a machine learning model for blood oxygen parameters based on 18 features, including the time series average values of the left and right prefrontal lobes and 10 frequency domain features of the corresponding parameters. For each subject, WPCO analysis was performed on the bilateral frontal lobe time series of each parameter. After averaging the time-frequency amplitude matrix of WPCO analysis at the time

Table 3. Feature differences between these groups (♦ P < 0.05)

the difference between two groups	Frequency interval	V 0.0095-0.02	IV 0.02-0.06	III 0.06-0.15	II 0.15-0.40	I 0.40-2.00	Mean
hemorrhagic and control	RTOI	0.244	0.190	0.419	0.003 ♦	0.001 ♦	0.202
	LTOI	0.231	0.452	0.372	0.002 ♦	0.001 ♦	0.011 ♦
	RHb	0.254	0.024 ♦	0.138	0.002 ♦	0.000 ♦	0.014 ♦
	LHb	0.100	0.146	0.193	0.028 ♦	0.001 ♦	0.263
	RHbO	0.012 ♦	0.018 ♦	0.287	0.022 ♦	0.157	0.284
	LHbO	0.001 ♦	0.059	0.378	0.035 ♦	0.274	0.214
	wTOI	0.001 ♦	0.000 ♦	0.000 ♦	0.000 ♦	0.007 ♦	0.006 ♦
	wHb	0.132	0.192	0.165	0.215	0.279	0.378
	wHbO	0.006 ♦	0.001 ♦	0.013 ♦	0.010 ♦	0.000 ♦	0.000 ♦
ischemic and control	RTOI	0.479	0.006 ♦	0.000 ♦	0.000 ♦	0.103	0.000 ♦
	LTOI	0.141	0.006 ♦	0.000 ♦	0.000 ♦	0.434	0.000 ♦
	RHb	0.363	0.060	0.000 ♦	0.010 ♦	0.108	0.288
	LHb	0.493	0.472	0.000 ♦	0.071	0.333	0.202
	RHbO	0.270	0.026 ♦	0.011 ♦	0.050 ♦	0.016 ♦	0.194
	LHbO	0.117	0.136	0.016 ♦	0.060	0.041 ♦	0.328
	wTOI	0.335	0.135	0.000 ♦	0.000 ♦	0.482	0.004 ♦
	wHb	0.178	0.229	0.135	0.109	0.002 ♦	0.080
	wHbO	0.116	0.444	0.395	0.378	0.002 ♦	0.055
hemorrhagic and ischemic	RTOI	0.233	0.093	0.000 ♦	0.001 ♦	0.001 ♦	0.000 ♦
	LTOI	0.096	0.022 ♦	0.000 ♦	0.000 ♦	0.001 ♦	0.000 ♦
	RHb	0.397	0.360	0.001 ♦	0.000 ♦	0.001 ♦	0.130
	LHb	0.133	0.208	0.013 ♦	0.001 ♦	0.000 ♦	0.328
	RHbO	0.104	0.432	0.025 ♦	0.002 ♦	0.004 ♦	0.412
	LHbO	0.039 ♦	0.267	0.055	0.004 ♦	0.017 ♦	0.303
	wTOI	0.002 ♦	0.000 ♦	0.036 ♦	0.308	0.010 ♦	0.195
	wHb	0.019 ♦	0.056	0.492	0.010 ♦	0.000 ♦	0.200
	wHbO	0.006 ♦	0.000 ♦	0.008 ♦	0.006 ♦	0.398	0.000 ♦

domain, a frequency sequence was obtained, which is divided into five bands that are the same as WT signals. Six features were extracted, that is, the average frequency of each frequency band and the average of the whole frequency sequence in five frequency bands. Hence, 18 features were extracted from each parameter to train the classification model. The process of feature screening, model training, and feature analysis is shown in Fig. 2.

Data preprocessing includes standardization and elimination of singular values. Standardization refers to scaling data to a specific small interval. After time-frequency domain analysis of blood oxygen parameters, the order of magnitude of frequency features between each physiological frequency band is completely different. If the original value is directly used for analysis, indicators with a higher order of magnitude will have a significant impact on the construction of the model, and indicators with a lower order of magnitude will be ignored. Here, we standardized the data in the training sample set of the machine learning model to ensure that the models for diagnosing cerebral hemorrhage and cerebral infarction have better classification accuracy and generalization performance so that data with different features have similar scales.

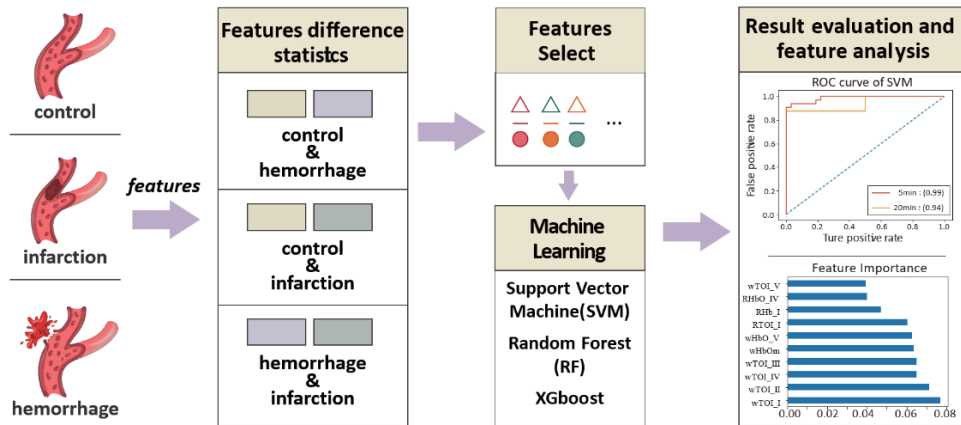


Fig. 2. Feature analysis and classification. First, the differences were analyzed between two groups of three subjects' features, and then, the features with significant differences were selected to train the machine learning classification model. Finally, we evaluated the trained model and analyzed the importance of features.

Based on the average values and standard deviations of features extracted from all subjects, Z score standardization was used to standardize each feature extracted from time-frequency analysis (Eq. (1)):

$$x = \bar{x} / \text{std}(x) \quad (1)$$

where x , \bar{x} and $\text{std}(x)$ represent one eigenvalue, standardized eigenvalue and standard deviation of x extracted from all subjects, respectively. This method is applicable to cases in which the maximum and minimum values of data are unknown, or there is a singular value. In the next study, the experimental data of the new sample will be processed according to the existing standard deviation. The feature values obtained before and after standardization are in a similar range. The data quality improved without changing the numerical ordering. The tail percentile removal method was used to eliminate the singular value, with 97.5% and 2.5% as the upper and lower limit critical values, respectively, to retain the maximum amount of information in the sample and ensure the classification result of the machine learning model.

The model was constructed with Datspell software, and the Python language version is 3.8. SVM, RF and XGBoost were selected as three machine learning models, which are provided by the scikit learning package in the Python environment. The method of cross validation was used to partition the dataset, avoiding the impact of the randomness of a single sample dataset partition on the model training error and generalization error evaluation, which is helpful to select a general robust parameter combination. The grid search strategy and traversal method were used to determine the optimal model parameters. The initial model parameters were identified by the convex optimization principle. According to the model performance and parameter adjustment results in the previous step, we gradually reduced the upper and lower limit range and step size until the approximate optimal model parameters were determined. The process of model training involved importing sample data, using the tenfold cross-validation as the training method, and determining the optimal parameters based on grid search.

The features of patients with cerebral hemorrhage, patients with cerebral infarction, and healthy controls were studied in pairs for classification model training and feature analysis. The three pairs were cerebral hemorrhage and healthy controls, cerebral infarction and healthy controls, and cerebral hemorrhage and cerebral infarction. Then, the features of cerebral hemorrhage patients, cerebral infarction patients, and healthy controls were used to train a three-classification model, the features of which were analyzed.

Accuracy, precision, recall rate, and *F1* score (Eq. (2)) were used to evaluate the performance of the model in this study. They were calculated based on the confusion matrix:

$$\begin{aligned}
 \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
 \text{Precision} &= \frac{TP}{TP + FP} \\
 \text{Recall} &= \frac{TP}{TP + FN} \\
 F1 &= \frac{2TP}{2TP + FP + FN}
 \end{aligned} \tag{2}$$

where *TP*, *TN*, *FP* and *FN* represent the number of correctly classified positive samples, correctly classified negative samples, wrongly classified negative samples and wrongly classified positive samples, respectively. Accuracy is the percentage of correctly classified samples in the total samples and is the most basic evaluation index in classification problems. It measures the performance of the model to a certain extent. When the proportion of sample categories is unbalanced, or there are extremely biased samples, the accuracy cannot be objectively evaluated. Precision, also known as the precision ratio, refers to the proportion of positive samples among all positive samples, and it measures the reliability of the model when classifying positive samples. The recall rate refers to the proportion of positive samples correctly classified as positive, and it measures the overall ability of the model to classify positive samples. Receiver operating characteristic (ROC) curve analysis reliably reflects the generalization performance of the model. Finally, we compare the reliability of the model by using Leave-One-Out cross-validation (LOOCV) with the tenfold cross-validation.

2.5. Data augmentation

Due to the relatively small amount of data, to strengthen the robustness of the training model, the 20-minute continuous signal (NDA) obtained from each subject was divided into four 5-minute-segment signals (DA); hence, the data were augmented four times. The same features were chosen for classification model training and feature importance analysis.

3. Results

3.1. Two classification results

We analyzed the difference in brain oxygen saturation signals between patients with cerebral hemorrhage and normal people. Based on the *t* test results in Table 3, we selected features with differences less than 0.05 for model training. We trained the classifier using individual TOI, Hb, and HbO features, and then combined the three series of features to train the classifier. Finally, we added the features calculated through WPCO to train the classifier and compared the classification results. The selected features are highlighted in bold in Table 3. We trained SVM, RF, and XGBoost classifiers separately. To avoid result redundancy, we demonstrate the best SVM classification results in Figs. 3(a-f). The classification results of patients with cerebral ischemia and the normal control group were also the best with the SVM classifier, as shown in Figs. 3(g-l). The classification results of hemorrhagic stroke and ischemic stroke were also the best for SVM classifiers, as shown in Figs. 3(m-r). Figure 4(a) shows the ROC curves of the three classification models before and after data augmentation. Figure 4(b) shows the ROC curves of the three classification models before and after data expansion. Figure 4(c) shows the ROC curves of the three classification models before and after data expansion.

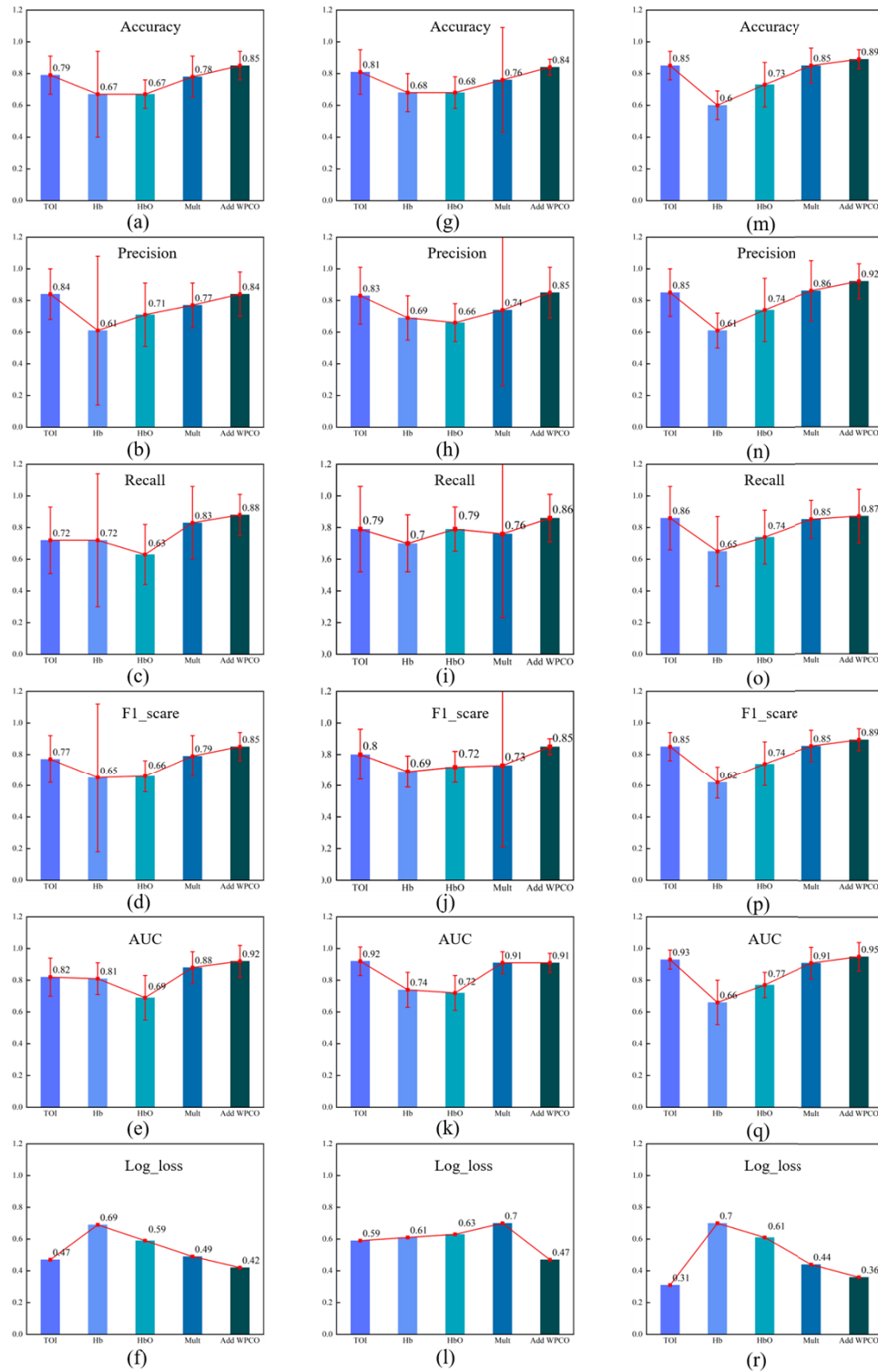


Fig. 3. The comparison of various evaluation parameters corresponding to different data combinations of the SVM model. The first column of the subfigures (a-f) corresponds to hemorrhagic stroke and controls, the second column corresponds (g-l) to ischemic stroke and controls, and the last column (m-r) corresponds to hemorrhagic stroke or ischemic stroke.

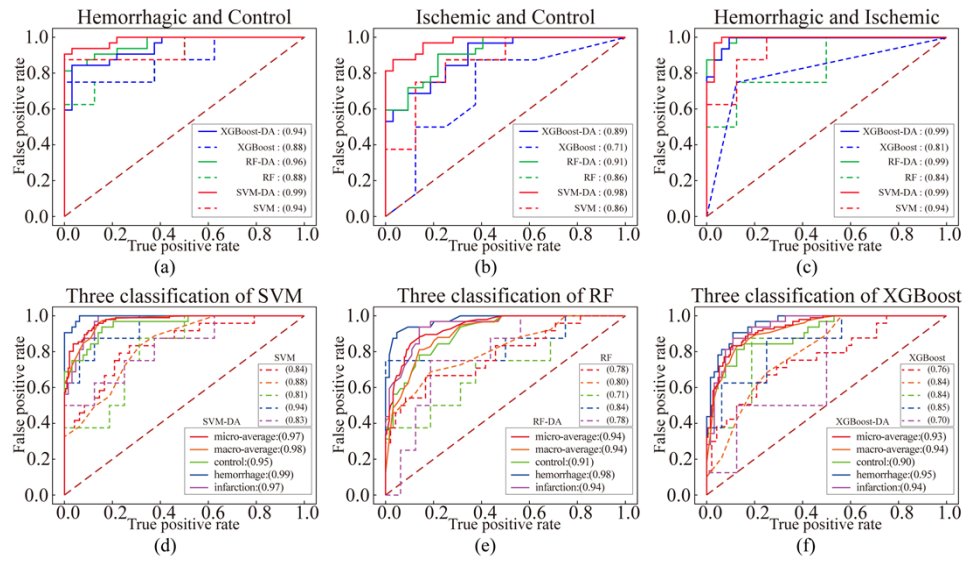


Fig. 4. ROC curve analysis: (a) ROC curve before and after data augmentation of three classifiers for hemorrhagic stroke and control group; (b) the ROC curves of three classifiers in the ischemic stroke and control group before and after data enhancement; (c) the ROC curves of the three classifiers for hemorrhagic stroke and ischemic stroke before and after data enhancement; (d)-(f) the ROC curves before and after data augmentation for three classifiers corresponding to hemorrhagic stroke, ischemic stroke, and the control group.

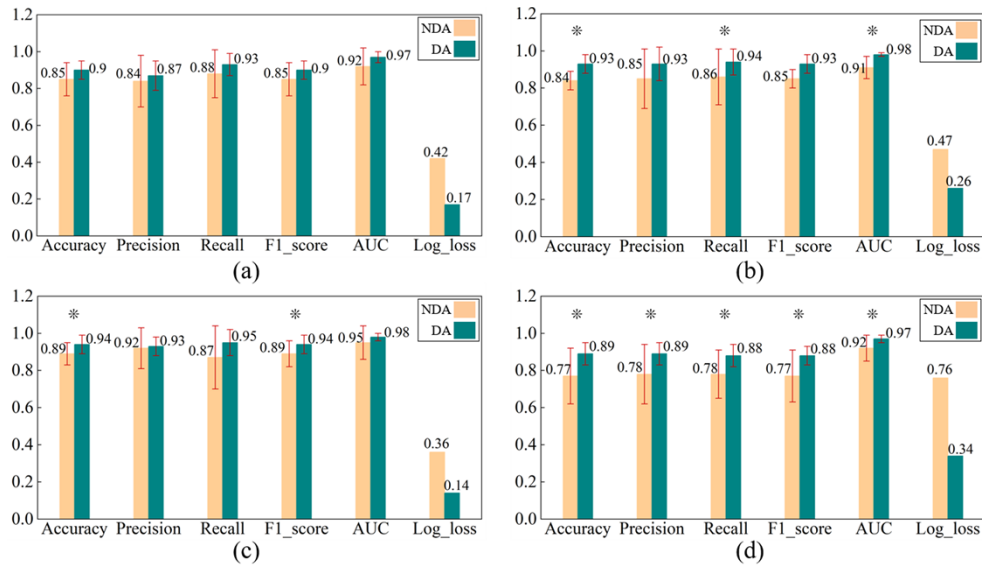


Fig. 5. Comparison of results before (NDA) and after data augmentation (DA). (a) shows the evaluation results of the SVM classifier before and after data expansion for hemorrhagic stroke and the normal control group; (b) shows the evaluation results of the SVM classifier before and after data expansion for ischemic stroke and normal control group; (c) shows the evaluation results of the SVM classifier before and after data expansion for hemorrhagic stroke and ischemic stroke; (d) shows the evaluation results of the SVM classifiers before and after data expansion for hemorrhagic stroke, ischemic stroke, and the normal control group. (* $P < 0.05$)

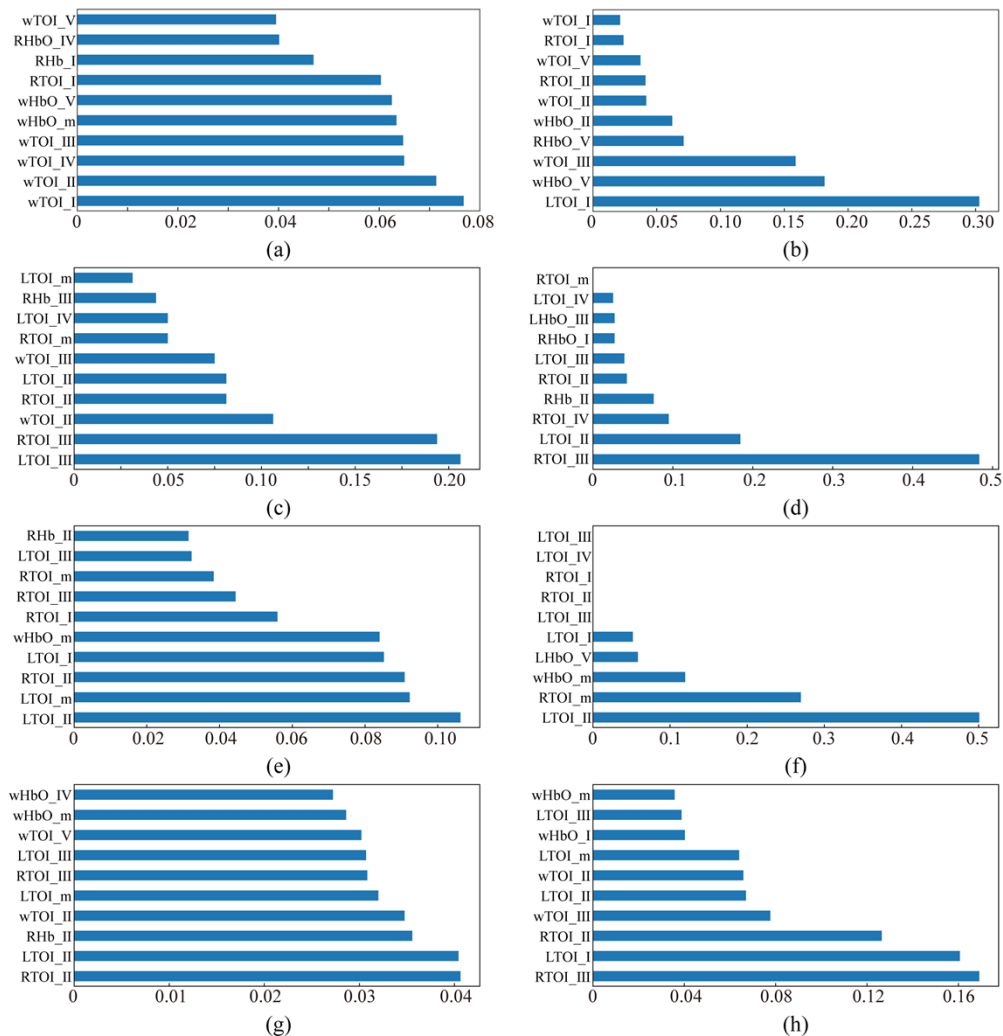


Fig. 6. Feature importance. (a) and (b) show the top ten feature rankings of the RF and XGBoost classification models for hemorrhagic stroke, respectively, compared to the control group; (c) and (d) show the top ten feature rankings of the RF and XGBoost classification models for ischemic stroke, respectively, and the control groups; (e) and (f) show the top ten feature rankings of the RF and XGBoost classification models for hemorrhagic stroke and ischemic stroke, respectively; (g) and (h) show the top ten feature rankings of the RF and XGBoost classification models for hemorrhagic stroke and ischemic stroke, respectively, and the control groups.

3.2. Three-class classification results

When training a three-class classifier, all the features used in the previous section are selected.

The classification results of the three classifiers were not significantly different, but the SVM classifier still performed better. We show the classification results in Fig. 5(d) in the next section. Figures 4(d-f) shows the ROC curves of the SVM, RF, and XGBoost classifiers, respectively.

3.3. Data augmentation results

The training results of the classifier before data augmentation (NDA) and after data augmentation (DA) are shown in Fig. 5. Figure 5(a) shows the classification results of hemorrhagic stroke and the control group, Fig. 5(b) shows the classification results of ischemic stroke and the control group, Fig. 5(c) shows the classification results of hemorrhagic stroke and ischemic stroke, and Fig. 5(d) shows the classification results of hemorrhagic stroke, ischemic stroke and the control group. The results of SVM, RF, and XGBoost classifiers are similar, Fig. 5 shows all the results of the SVM classifier. The LOOCV result of the binary-classification model is 90%, which is the same as the result of cross validation with 10-fold CV. The LOOCV result of the three-classification model is 90%, while the result of 10-fold CV is 89%.

3.4. Result of feature importance analysis

We used three classifiers – SVM, RF, and XGboost – to train the classification model. Among them, SVM had the best classification performance, but the three classifiers had significant differences in classification performance. Because the classifier training in this experiment was implemented using the scikit-learn library in the Python language and the SVM classifier used the Gaussian kernel function, which does not support feature importance, we used RF and XGBoost classifiers for feature importance analysis. The ranking results of feature importance are shown in Fig. 6, and Figs. 6(a) and (b) correspond to the feature importance analysis results of the RF and XGBoost classifiers in the classification of hemorrhagic stroke and the control group, respectively; Figs. 6(c) and (d) correspond to the feature importance analysis results of the RF and XGBoost classifiers in the classification of ischemic stroke and the control group, respectively; Figs. 6(e) and (f) correspond to the feature importance analysis results of the RF and XGBoost classifiers in the classification of hemorrhagic stroke and ischemic stroke, respectively; and Figs. 6(g) and (h) correspond to the feature importance analysis results of the RF and XGBoost classifiers in the classification of hemorrhagic stroke, ischemic stroke, and the normal control group, respectively. Each classifier is trained with more than forty features. Figure 6 shows the top ten features in importance ranking for each classifier. We can see that, in the RF classifiers, the importance of each feature is relatively balanced. However, in XGboost classifiers, there is a significant difference in feature importance. In summary, TOI_II and TOI_I contribute the most to the classifier.

4. Discussion

We divided the cerebral hemorrhage, cerebral infarction and healthy control groups into pairs to train the classification model and analyzed the classification results. According to the model evaluation results, and from the perspective of different types of signal parameters, the main difference among the groups was the TOI signal. For each group, after adding the cross-correlation signal features of the left and right frontal lobes, all evaluation indicators of the model improved, indicating that the left and right frontal lobe connectivity differed among the groups, indicating that hemorrhage or infarction might occur on one side. By ranking the importance of features, it was found, that between the cerebral hemorrhage group and the healthy control group, the most significant feature was the cross-correlation feature of TOI in the left and right frontal lobes, among which the feature corresponded to myogenic activity, respiratory activity and cardiac activity, involving the difference between cardiac pumping and respiratory activity [38]. The most important representation shows heart activity. Between the cerebral infarction group and the healthy control group, the most important feature was the TOI signal in the third frequency band. This difference mainly corresponds to the myogenic activity in physiological activities, specifically reflected in the tense contraction of vascular smooth muscle. However, most cases of cerebral infarction are caused by cerebral vascular atherosclerosis, in line with this fact [38]. Between the

cerebral hemorrhage group and the cerebral infarction group, the feature importance of the TOI signal in the second frequency band was the largest, corresponding to respiratory activity [38] and indicating that the respiratory activities of patients with cerebral hemorrhage and patients with cerebral infarction are significantly different. Finally, we constructed a three-classification model, and the accuracy of the model results was greater than 85%, with effective reference significance for clinical practice.

The classification model of patients with cerebral hemorrhage, patients with cerebral infarction and healthy humans built by us has an accuracy rate of 89% (three-classification model) after data augmentation. The model has the advantages of convenience, immediacy and high accuracy. Its program can be embedded in small equipment to facilitate carrying for use in limited conditions. In case of emergencies, it is only necessary to collect a 5-min cerebral oxygen signal of a patient to obtain the classification results in a few seconds, providing doctors with important guidance information in a timely manner, which is of great significance for patients' diagnosis and further timely and effective treatment. The minimum accuracy rate of the model was greater than 85%, and the accuracy rate of the model trained after data augmentation was greater than 90% (among in binary-classification model). For example, in the field of signal processing, especially in enhancing classification data from EEG signals, researchers have reported similar improvements in classification accuracy through data augmentation. [47] The accuracy is at a high level compared with that from similar research [48,28]. It has high reliability but has some room for improvement. The results show that this small-volume detection instrument can also make reliable diagnoses for patients with cerebral hemorrhage and cerebral infarction using near-infrared spectroscopy technique. Because of its simple operation and portability, it is of great significance in the screening for stroke diseases in a large population and for improving the diagnosis of cerebral hemorrhage symptoms using brain CT in most current diagnostic models. This automatic classification framework shows the potential to shorten the diagnosis time before determining the treatment mode of patients [49], which is of great significance for improving the treatment effects of patients and reducing the degree of sequelae in patients.

We utilized the tenfold cross-validation method, which inherently evaluates the model's performance on independent subsets of the data. The dataset was randomly partitioned into ten equal parts, and the model was trained and tested ten times, each time using a different partition as the validation set, while the remaining nine partitions served as the training set. This approach ensures that the model is evaluated based on diverse and independent subsets of the data, providing a rigorous assessment of its generalizability and performance. Although the tenfold cross-validation method ensures robustness to some extent, it is important to note that the performance of the model heavily relies on the quality and representativeness of the training data. If the training dataset is biased, incomplete, or insufficient in capturing the true variations in clinical cases, the model's reliability could be compromised. Additionally, the model's performance might vary across different populations or settings, which can further impact its generalizability and reliability in practical applications. Therefore, further research and refinement are necessary to address these limitations and enhance the model's applicability and accuracy in real-world clinical settings.

In our study, we have addressed some limitations of fNIRS: 1) the spatially resolved spectroscopy algorithm was used to calculate the absolute value of tissue oxygen saturation without contamination from the scalp; 2) spontaneous LFO reflects the hemodynamic changes of whole cerebral blood vessels, which is not restricted to the detection area or the shallow depth of the brain; and 3) with the help of machine learning, it is possible to find a determined way to recognize hemorrhagic stroke and ischemic stroke. It is of great significance to recognize stroke in time with low cost in a readily available manner. Based on existing data we verified our method in stroke recognition. In future studies, more subjects will be included to build a more robust

model. With sufficient data, it is possible to introduce methods such as deep learning to improve the accuracy and performance of the model.

5. Conclusion

In this study, a recognition framework for cerebral stroke is proposed based on feature analysis and machine learning using fNIRS signals. Three machine learning methods were established to classify patients with cerebral hemorrhage, patients with cerebral infarction, and healthy controls. The importance of important features in classification was evaluated. The reasons for the significant differences among the groups were analyzed. The features extracted from TOI parameters contribute the most importance to the classification of patients with cerebral hemorrhage and cerebral infarction. Effective classification results based on the signals obtained from short-term sampling show that we have developed a potential solution for noninvasive and quick identification of cerebrovascular conditions in patients with cerebral hemorrhage and cerebral infarction.

Funding. National Natural Science Foundation of China (81901907, 82172112); Beijing Institute of Technology Research Fund Program for Young Scholars; Fundamental Research Funds for the Central Universities (LY2022-22).

Acknowledgments. Authors also thank the Analysis and Testing Center of Beijing Institute of Technology.

Disclosures. The authors declare no conflicts of interest.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. J. Zhao and R. Liu, "Stroke 1-2-0: a rapid response programme for stroke in China," *Lancet Neurol.* **16**(1), 27–28 (2017).
2. V. L. Feigin, M. Brainin, B. Norrving, S. Martins, R. L. Sacco, W. Hacke, M. Fisher, J. Pandian, and P. Lindsay, "World Stroke Organization (WSO): Global stroke fact sheet 2022," *International Journal of Stroke* **17**(1), 18–29 (2022).
3. C. W. Tsao, A. W. Aday, and Z. I. Almarzooq, *et al.*, "Heart Disease and Stroke Statistics-2022 update: a report from the American Heart Association," *Circulation* **145**(8), E153–E639 (2022).
4. L. Davies and C. Delcourt, "Current approach to acute stroke management," *Intern. Med. J.* **51**(4), 481–487 (2021).
5. E. O. Terecoasa, R. A. Radu, A. Negrila, I. Enache, B. Casaru, and C. Tiu, "Pre-hospital delay in acute ischemic stroke care: current findings and future perspectives in a tertiary stroke center from romania-a cross-sectional study," *Medicina* **58**(8), 1003 (2022).
6. C. Provost, M. Soudant, L. Legrand, W. Ben Hassen, Y. Xie, S. Soize, R. Bourcier, J. Benzakoun, M. Edjlali, G. Boulouis, H. Raoult, F. Guillemin, O. Naggara, S. Bracard, and C. Oppenheim, "Magnetic resonance imaging or computed tomography before treatment in acute ischemic stroke effect on workflow and functional outcome," *Stroke* **50**(3), 659–664 (2019).
7. S. Wu, B. Wu, and M. Liu, *et al.*, "Stroke in China: advances and challenges in epidemiology, prevention, and management," *Lancet Neurol.* **18**(4), 394–405 (2019).
8. A. Garcia-Tornel, L. Sero, and X. Urrea, *et al.*, "Workflow times and outcomes in patients triaged for a suspected severe stroke," *Ann. Neurol.* **92**(6), 931–942 (2022).
9. J. J. Park, Y. Kim, C. L. Chai, and J. P. Jeon, "Application of near-infrared spectroscopy for the detection of delayed cerebral ischemia in poor-grade subarachnoid hemorrhage," *Neurocrit. Care* **35**(3), 767–774 (2021).
10. G. A. Donnan, M. Fisher, M. Macleod, and S. M. Davis, "Stroke," *Lancet* **371**(9624), 1612–1623 (2008).
11. H. Ayaz, M. Izzetoglu, K. Izzetoglu, B. Onaral, and B. Ben Dor, "Early diagnosis of traumatic intracranial hematomas," *J. Biomed. Opt.* **24**(05), 1–10 (2019).
12. C. S. Robertson, E. L. Zager, R. K. Narayan, N. Handly, A. Sharma, D. F. Hanley, H. Garza, E. Maloney-Wilensky, J. M. Plaum, C. H. Koenig, A. Johnson, and T. Morgan, "Clinical evaluation of a portable near-infrared device for detection of traumatic intracranial hematomas," *Journal of Neurotrauma* **27**(9), 1597–1604 (2010).
13. V. Kontogiannis, I. Hostettler, R. J. Brogan, M. Raza, A. Harper-Payne, H. Kareem, M. Boutelle, and M. Wilson, "Detection of intracranial hematomas in the emergency department using near infrared spectroscopy," *Brain Injury* **33**(7), 875–883 (2019).
14. Y. C. Liu, Y. R. Yang, Y. A. Tsai, R. Y. Wang, and C. F. Lu, "Brain activation and gait alteration during cognitive and motor dual task walking in stroke-a functional near-infrared spectroscopy study," *IEEE Trans. Neural Syst. Rehabil. Eng.* **26**(12), 2416–2423 (2018).
15. G. Giacalone, M. Zanoletti, R. Re, B. Germinario, D. Contini, L. Spinelli, A. Torricelli, and L. Roveri, "Time-domain near-infrared spectroscopy in acute ischemic stroke patients," *Neurophotonics* **6**(01), 1 (2019).

16. R. Hiramatsu, M. Furuse, R. Yagi, H. Ohnishi, N. Ikeda, N. Nonoguchi, S. Kawabata, S. Miyachi, and T. Kuroiwa, "Limit of intraoperative near-infrared spectroscopy monitoring during endovascular thrombectomy in acute ischemic stroke," *Interv Neuroradiol* **24**(1), 57–63 (2018).
17. A. Gereg, S. Wojtkiewicz, P. Sawosz, M. Kacprzak, B. Toczyłowska, K. Bejm, F. Skibniewski, A. Sobotnicki, A. Gacek, R. Maniewski, and A. Liebert, "Assessment of the brain ischemia during orthostatic stress and lower body negative pressure in air force pilots by near-infrared spectroscopy," *Biomed. Opt. Express* **11**(2), 1043–1060 (2020).
18. Y.-H. Chen and M. Sawa, "Trends and challenges of wearable multimodal technologies for stroke risk prediction," *Sensors* **21**(2), 460 (2021).
19. T. Gao, C. Zou, J. Li, C. Han, H. Zhang, Y. Li, X. Tang, and Y. Fan, "Identification of moyamoya disease based on cerebral oxygen saturation signals using machine learning methods," *J. Biophotonics* **15**(7), e202100388 (2022).
20. Y. Li, Y. Wang, Y. Li, J. Wang, L. Li, and Zhang, "Wavelet analysis of cerebral oxygenation signal measured by near infrared spectroscopy in subjects with cerebral infarction," *Microvasc. Res.* **80**(1), 142–147 (2010).
21. J. B. M. Zeller, A. Katzorke, L. D. Mueller, J. Breunig, F. B. Haeussinger, J. Deckert, B. Warrings, M. Lauer, T. Polak, and M. J. Herrmann, "Reduced spontaneous low frequency oscillations as measured with functional near-infrared spectroscopy in mild cognitive impairment," *Brain Imaging and Behavior* **13**(1), 283–292 (2019).
22. S. Becker, F. Klein, K. Koenig, C. Mathys, T. Liman, and K. Witt, "Assessment of dynamic cerebral autoregulation in near-infrared spectroscopy using short channels: A feasibility study in acute ischemic stroke patients," *Front. Neurol.* **13**, 1028864 (2022).
23. H. W. Schytz, F. M. Amin, J. Selb, and D. A. Boas, "Non-invasive methods for measuring vascular changes in neurovascular headaches," *J Cereb Blood Flow Metab* **39**(4), 633–649 (2019).
24. G. Biccato, G. Narula, G. Brandi, A. Eisele, S. Schulthess, S. Friedl, J. F. Willms, L. Westphal, and E. Keller, "Functional NIRS to detect covert consciousness in neurocritical patients," *Clin. Neurophysiol.* **144**, 72–82 (2022).
25. G. Biccato, E. Keller, M. Wolf, G. Brandi, S. Schulthess, S. G. Friedl, J. F. Willms, and G. Narula, "Increase in low-frequency oscillations in fNIRS as cerebral response to auditory stimulation with familiar music," *Brain Sci.* **12**(1), 42 (2021).
26. H. Ren, X. Jiang, K. Xu, C. Chen, Y. Yuan, C. Dai, and W. Chen, "A review of cerebral hemodynamics during sleep using near-infrared spectroscopy," *Front. Neurol.* **11**, 524009 (2020).
27. M. K. Yeung and A. S. Chan, "A systematic review of the application of functional near-infrared spectroscopy to the study of cerebral hemodynamics in healthy aging," *Neuropsychology Review* **31**(1), 139–166 (2021).
28. M. S. Sirsat, E. Fermé, and J. Câmara, "Machine learning for brain stroke: a review," *Journal of Stroke & Cerebrovascular Diseases* **29**(10), 105162 (2020).
29. V. V. Nair, B. R. Kish, H. C. Yang, Z. Y. Yu, H. Guo, Y. J. Tong, and Z. H. Liang, "Monitoring anesthesia using simultaneous functional near infrared spectroscopy and electroencephalography," *Clin. Neurophysiol.* **132**(7), 1636–1646 (2021).
30. J. A. Zhang, Y. Wang, Y. Zhang, B. Li, and Y. Zhang, "Enhanced written vs. verbal recall accuracy associated with greater prefrontal activation: A near-infrared spectroscopy study," *Front. Behav. Neurosci.* **15**, 601698 (2021).
31. R. Li, X. X. Ye, G. P. Li, X. K. Cao, Y. X. Zou, S. H. Yao, F. Luo, L. Zhang, and W. B. Dong, "Effects of different body positions and head elevation angles on regional cerebral oxygen saturation in premature infants of China," *Journal of Pediatric Nursing* **55**, 1–5 (2020).
32. L. Shi, J. F. Xu, J. G. Wang, M. H. Zhang, F. Liu, Z. U. Khan, S. Y. Liu, W. Zhou, A. Y. Qian, J. G. Zhang, and M. Zhang, "Automated pupillometry helps monitor the efficacy of cardiopulmonary resuscitation and predict return of spontaneous circulation," *American Journal of Emergency Medicine* **49**, 360–366 (2021).
33. X. Yang, X. P. Lei, L. Y. Zhang, L. P. Zhang, and W. B. Dong, "The application of near-infrared spectroscopy in oxygen therapy for premature infants," *J. Matern.-Fetal Neonat. Med.* **33**(2), 283–288 (2020).
34. B. Ze, L. L. Liu, G. S. Y. Jin, M. N. Shan, Y. H. Geng, C. L. Zhou, T. Q. Wu, H. Wu, and X. L. Hou, "Near-infrared spectroscopy monitoring of cerebral oxygenation and influencing factors in neonates from high-altitude areas," *Neonatology* **118**(3), 348–353 (2021).
35. Y. Teng, H. Ding, Q. Gong, Z. Jia, and L. Huang, "Monitoring cerebral oxygen saturation during cardiopulmonary bypass using near-infrared spectroscopy: the relationships with body temperature and perfusion rate," *J. Biomed. Opt.* **11**(2), 024016 (2006).
36. B. W. Hyndman, R. I. Kitney, and B. M. Sayers, "Spontaneous rhythms in physiological control systems," *Nature* **233**(5318), 339–341 (1971).
37. S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. C. Barger, and R. J. Cohen, "Power spectrum analysis of heart-rate fluctuation - a quantitative probe of beat-to-beat cardiovascular control," *Science* **213**(4504), 220–222 (1981).
38. Y.-K. Jan, S. Shen, R. D. Foreman, and W. J. Ennis, "Skin blood flow response to locally applied mechanical and thermal stresses in the diabetic foot," *Microvasc. Res.* **89**, 40–46 (2013).
39. F.-L. Wu, W. T. Wang, F. Liao, Y. Liu, J. Li, and Y.-K. Jan, "Microvascular control mechanism of the plantar foot in response to different walking speeds and durations: implication for the prevention of foot ulcers," *International Journal of Lower Extremity Wounds* **20**(4), 327–336 (2021).
40. W. T. Tseng, W. C. Shann, B. C. Chen, Y. C. Chang, and M. L. Tsai, "Sympathetic nerve activity-induced blood pressure fluctuations are stacked on linearly: A simulation study," *Neurophysiology* **50**(4), 243–248 (2018).

41. Z. Li, M. Zhang, Q. Xin, S. Luo, W. Zhou, R. Cui, and L. Lu, "Assessment of cerebral oxygenation oscillations in subjects with hypertension," *Microvasc. Res.* **88**, 32–41 (2013).
42. Q. Han, Z. Li, Y. Gao, W. Li, Q. Xin, Q. Tan, M. Zhang, and Y. Zhang, "Phase synchronization analysis of prefrontal tissue oxyhemoglobin oscillations in elderly subjects with cerebral infarction," *Med. Phys.* **41**(10), 102702 (2014).
43. Q. Han, M. Zhang, W. Li, Y. Gao, Q. Xin, Y. Wang, and Z. Li, "Wavelet coherence analysis of prefrontal tissue oxyhaemoglobin signals as measured using near-infrared spectroscopy in elderly subjects with cerebral infarction," *Microvasc. Res.* **95**, 108–115 (2014).
44. Z. Li, M. Zhang, R. Cui, Q. Xin, L. Liqian, W. Zhou, Q. Han, and Y. Gao, "Wavelet coherence analysis of prefrontal oxygenation signals in elderly subjects with hypertension," *Physiol. Meas.* **35**(5), 777–791 (2014).
45. Z. Li, M. Zhang, Q. Xin, G. Chen, F. Liu, and J. Li, "Spectral analysis of near-infrared spectroscopy signals measured from prefrontal lobe in subjects at risk for stroke," *Med. Phys.* **39**(4), 2179–2185 (2012).
46. Z. Li, M. Zhang, Q. Xin, S. Luo, R. Cui, W. Zhou, and L. Lu, "Age-related changes in spontaneous oscillations assessed by wavelet transform of cerebral oxygenation and arterial blood pressure signals," *J. Cereb. Blood Flow Metab.* **33**(5), 692–699 (2013).
47. M. Mikolajczyk and Grochowski, "Data augmentation for improving deep learning in image classification problem," presented at the *2018 International Interdisciplinary PhD Workshop (IIPhDW)* 2018.
48. S. Mainali, M. E. Darsie, and K. S. Smetana, "Machine learning in action: stroke diagnosis and outcome prediction," *Front. Neurol.* **12**, 734345 (2021).
49. D. Antipova, L. Eadie, A. Macaden, and P. Wilson, "Diagnostic accuracy of clinical tools for assessment of acute stroke: a systematic review," *BMC Emerg. Med.* **19**(1), 49 (2019).